DOCUMENT RESUME

ED 194 577 TM 800 706

AUTHOR Ward, Barbara
TITLE The National Assessment Approach to Objectives and Exercise Development.
INSTITUTION Education Commission of the States, Denver, Colo. National Assessment of Educational Progress.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO NAEP-12-IP-55
PUB DATE Sep 80
GRANT NIE-G-80-0003
NOTE 50p.: For related document see ED 067 402.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests: *Behavioral Objectives: *Educational Assessment: Elementary Secondary Education: Evaluation Criteria: Federal Programs: *National Surveys: Program Costs: *Program Development: Scoring: Test Bias: Test Construction: Testing: Testing Problems: *Test Items
IDENTIFIERS National Assessment of Educational Progress

ABSTRACT

The National Assessment of Educational Progress (NAEP) item development procedures, possible improvements or alternatives to these procedures, and methods used to control potential sources of errors of interpretation are described. Current procedures call for the assessment of 9-, 13- and 17-year-olds in subject areas typically taught in schools. Objectives and items are developed and reviewed through a consensus process which involves subject-matter experts, educators and lay persons. Reviews help to establish the accuracy and content validity of items and to guard against potential item bias. Procedures used to prepare item booklets, administer the assessment and score the items are designed to reduce sources of error and to ensure nearly identical conditions across assessments so measurements of change are accurate. Possible alternative procedures, including structuring assessments by means other than subject areas, using an assessment framework other than objectives, developing objectives through textbook analysis, employing objectives developed by other organizations, using a contractor other than NAEP to monitor development and changing the number or type of reviewers, are also discussed. Procedures that would benefit from further research are identified. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED194577

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

THE NATIONAL ASSESSMENT APPROACH TO OBJECTIVES AND EXERCISE DEVELOPMENT

No. 12-IP-55

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
T. Pratt

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

by Barbara Ward

Assisted by the Staff of the National Assessment of Educational Progress

Education Commission of the States
Suite 700, 1860 Lincoln Street
Denver, Colorado 80295

TM 800 709

September 1980

The National Assessment of Educational Progress is funded by the National Institute of Education. It is under contract with the Education Commission of the States. It is the policy of the Education Commission of the States to take affirmative action to prevent discrimination in its policies, programs and employment practices.

The National Assessment of Educational Progress is an education research project mandated by Congress to collect and report data, over time, on the performance of young Americans in various learning areas. National Assessment makes available information on assessment procedures and materials to state and local education agencies and others.

The work upon which this publication is based was performed pursuant to Grant No. NIE-G-80-0003 of the National Institute of Education. It does not, however, necessarily reflect the views of that agency.

TABLE OF CONTENTS

PREFACE 1

INTRODUCTION 3
The Developmental Process 3
Criteria for Assessment Items 4
Sources of Error in Interpreting National Assessment Data 5

CHAPTER 1 Selection of Assessment Areas and Development of Objectives 9
Selection of Areas To Be Assessed 9
Objectives Development 11
Area Specifications 17

CHAPTER 2 Development of Assessment Items and Background Questions 21
Item Writing 22
Field Testing 25
Item Reviews 28
Item Selection 30
Background Questions 31

CHAPTER 3 Preparation of Item Booklets, Item Administration and Scoring 35
Preparation of Item Booklets 35
Administration Procedures 38
Scoring 39

CHAPTER 4 Summary 43

BIBLIOGRAPHY 49

NOV 17 1980

PREFACE

The following paper is one of seven required by the National Institute of Education (NIE) grant to the Education Commission of the States for the administration of the National Assessment of Educational Progress (NAEP). The papers are intended to raise significant questions about NAEP functions and procedures in order to apprise the Assessment Policy Committee of its technical and administrative options as NAEP moves into a second decade of activity. According to the NIE Program Announcement, each paper should include "discussions about the significance of a national assessment of educational progress for (a) informing public opinion, (b) influencing educational and other social policies, (c) identifying specific educational and research needs and (d) contributing to the goal of improving education." Thus, the papers combine broad theoretical considerations with very concrete discussions of potential options and their trade offs. All the papers are suggestive, rather than definitive -- starting points for discussion rather than end points.

Two fundamental questions undergird these papers. First, if one were to design a national assessment program from scratch to serve the needs of the eighties and nineties, what would it look like? What theory would it draw upon? What constituents would it serve? How would it operate? And second, given that an assessment has been operating for 10 years, how do answers to these questions match the current NAEP? How could the current NAEP be changed to accommodate needs and technology unforeseen when it was begun in the 1960s? As these questions weave in and out of the discussions of various issues, it will become clear that the answers one gives will probably depend upon one's assumptions about such things as the way society, education and science work, as well as one's familiarity with the theory, politics, history and operation of the current National Assessment.

There are many ways to implement the 1978 legislation mandating a national assessment. Some are undoubtedly better than others in certain respects; all

have both advantages and drawbacks, complicating even the simplest comparisons among alternatives. In such a situation, it is imperative that advocates for various models of a national assessment adopt an open, interdisciplinary approach to rethinking the Assessment, listening carefully to each other and respecting professional differences of opinion.

Individuals charged with making policy decisions about NAEP -- members of the Assessment Policy Committee and members of the U.S. Congress -- will find that, in many instances, attempts to strengthen the Assessment in one regard will weaken it in some other regard. Given a certain level of funding, advances in some areas come at the expense of others. The National Assessment can neither be all things to all people nor a single, unchanging enterprise. It will always represent a combination of various tensions. The question for its future is not how do we eliminate those tensions? Rather it is: How do we harness those tensions creatively to keep the Assessment abreast of the times and continually relevant to the needs of American education?

INTRODUCTION

The National Assessment of Educational Progress (NAEP) is a unique source of information about changes in American educational achievement over time. The instruments and procedures employed to gather this information directly affect the ways it can be used and interpreted. This paper discusses National Assessment's item development procedures, the relationship of these procedures to the project's goals and possible improvements or alternative procedures.

The major goals of the Assessment are to report the current status of American education and to detect changes in that status over time. To fulfill these goals, National Assessment (1) elicits the views of a broad cross-section of people involved in education so that results will be useful to diverse constituencies, (2) measures a wide range of knowledge and skills so that areas in which changes do occur can be identified and (3) attempts to make extremely accurate measurements so that small changes can be reported with confidence that they are not chance variations due to sampling.

Three age groups of students -- 9-, 13- and 17-year-olds -- are assessed in a number of subject areas; the subject areas are periodically reassessed to see if students of the same age are gaining or losing ground. Young adults aged 26-35 and 17-year-olds not attending school are assessed when resources permit.

The Developmental Process

In assessing a learning area, a framework for structuring item development is first established. This framework most often takes the form of objectives, although content by process matrices have been used for some subject areas. Whatever form the objectives for a learning area take, National Assessment procedures have traditionally required that subject-matter specialists see them as authentic for the subject, that educators believe them to be objectives of American education and that parents and other interested lay citizens agree

that they are important. The objectives are developed by representatives of different regions of the country, sizes and types of communities and minority groups.

After a framework has been created, exercises (items)[1] are developed. The items are then field tested to provide information on potential problems with vocabulary or instructions, approximate item difficulty and amount of time needed to administer an item. Items are reviewed by subject-matter experts and educators to insure that they are unambiguous and measure the objective they are intended to measure and by lay persons to determine whether they are free from bias and measure things considered important for young people to know and be able to do. Background questions, some of which are answered by assessment respondents and some by school officials, are developed to provide information about respondents' backgrounds and experiences with the subject being assessed.

Criteria for Assessment Items

National Assessment strives to meet a number of criteria in developing assessment items. An important goal of the Assessment is to measure changes in achievement over time, but it is impossible to predict with certainty areas in which changes will occur or areas that will become more or less important in the future. Thus, an assessment includes broad coverage of many aspects of a subject area so that changes that occur can be detected. Items must allow for continuity over time. They must not be so time-bound that they cannot be used in subsequent assessments to measure change, unless a time-bound item is the only appropriate way to measure an objective, or so oriented toward future trends that their present measurement does not provide meaningful information. To describe the range of tasks that different numbers of people can do, items are purposefully designed to have different difficulty levels: (1) items are included that are easy -- most students at an age level can do them, (2) that are of medium difficulty -- about half the students can achieve them, and (3) that are hard -- relatively few students are able to accomplish them.

---

[1]The terms item and exercise are used interchangeably.

Sources of Error in Interpreting National Assessment Data

In reporting assessment data, there is a need to provide information about how confident one can be about the conclusions drawn from the data. Although National Assessment does not draw conclusions in the same sense as one does when attempting to make inferences after testing hypotheses, National Assessment data are used to describe the situation that exists in American education and that description must be as close to "reality" as possible.

Traditionally, sources of error in interpreting results of individuals' test scores are described in terms of validity and reliability. Empirical evidence of validity and reliability has thus taken the form of coefficients arrived at by correlating the results of two testings of a group. High correlations are in part determined by the extent to which the two measures spread the scores of individuals. Since some items are included in the Assessment that nearly everyone can answer and some that very few can answer, variations among scores may not be maximized. Thus, traditional procedures for obtaining measures of reliability and validity often will not provide meaningful information for National Assessment data and are not of concern when describing group performance. However, the concepts of reliability and validity may be applicable to interpretations of National Assessment data.

Those responsible for National Assessment are highly concerned with identifying and minimizing potential sources of error in interpretations of the data. There is potential for error in all phases of the Assessment -- from objectives and item development to data collection to data reporting. For example, objectives may present an inaccurate picture of the subject area or items may not accurately represent the objectives. The context of the exercises -- their physical format, the mixture of items in a booklet, the length of the item booklet -- may be factors in student responses. Differences in exercise administration, student motivation, the physical setting of the assessment or exercise scoring also may affect performance. Some of the steps taken to control sources of error are discussed in this paper; others are addressed elsewhere in the issue paper series.

Two potential sources of error must be addressed in any interpretation of National Assessment results.

First, there is potential for error in the sample of the population used to estimate performance for that population. Second, there is potential for error in generalizing from performance on specific tasks included in an assessment to performance on similar tasks. Procedures used in collecting data may affect both types of errors. Measures taken to estimate population sampling errors are described in Issues in the Analysis and Analysis of Change of National Assessment Data, 1980.

Potential errors arising from the items assessed and the procedures used may affect the validity of Assessment data. Validity concerns the extent to which a test actually provides information on a particular interpretation or application it presumes to provide information on. As Cronbach described it, "Narrowly considered, validation is the process of examining the accuracy of a specific prediction or inference made from a test score.... More broadly, validation examines the soundness of all the interpretations of a test -- descriptive and explanatory interpretations as well as situation-bound predictions" (Cronbach, 1971, p. 443). Assessment results do not have validity in the abstract -- their validity must be determined with reference to the particular use to be made of the data.

Users of National Assessment data must carefully examine Assessment instruments and procedures to determine whether the data are valid for their purposes. National Assessment is designed to help users in making that decision through its attention to content validity. Content validity involves the systematic investigation of item content to determine whether items adequately cover the domain being measured. Such a validation procedure is commonly used in evaluating achievement tests. To help establish content validity, exercises are carefully reviewed by groups of consultants representing different backgrounds and interest groups. These reviewers inspect the items to determine whether they do in fact indicate achievement of the objectives they are designed to measure and whether the items measuring an objective are a sufficient indicator of achievement of the objective.

National Assessment items must also have a considerable degree of face validity -- that is, the assessment must "look meaningful" to those who respond to the items and to those who will use the results. Even though items may have content validity, if the

assessment does not appear to measure what it says it measures, it may be difficult to convince people that the results are in fact meaningful.

Concern for potential sources of errors in interpretation goes beyond the development of objectives and exercises to the procedures that are used in collecting and reporting data. Many of these procedures are detailed in other National Assessment publications. (See, for example, Procedural Handbook: 1977-78 Mathematics Assessment, 1979.) Procedures, such as using paced tapes to read items to students, using trained administrators to administer item booklets and having students mark their answers in the item booklets, were designed to reduce the effect of extraneous factors on achievement. Ongoing research into methodology is conducted in an attempt to determine the effect that various procedures have on performance.

The following chapters describe the procedures used by National Assessment in an effort to minimize the impact of various sources of error in developing objectives and exercises. Chapter 1 presents rationales for selecting areas to be assessed, methods for developing objectives and procedures for generating specific guidelines for exercise writers. Chapter 2 describes methods used for writing and field testing items, selecting items for the assessment and creating background questions. Chapter 3 discusses structuring of item booklets, administration of the assessment and scoring of items. Chapter 4 summarizes potential changes in procedures and trade offs that might be involved.

CHAPTER 1

SELECTION OF ASSESSMENT AREAS AND DEVELOPMENT OF OBJECTIVES

During the past 15 years, National Assessment staff have gained considerable experience in the development of objectives and exercises. They have found that while procedures can be standardized to a certain extent the unique characteristics of the learning area being assessed have an influence on the procedures used. For example, procedures appropriate for writing, an area with relatively few exercises, many of which are open-ended, may not be appropriate for mathematics, an area comprising many multiple-choice items. This paper emphasizes procedures common to all subject areas, noting some areas in which variations have occurred.

The scope and procedures of an assessment are necessarily impacted by cost factors. National Assessment places considerable emphasis on technical quality and is careful that while procedures are planned to be cost-efficient they still guarantee the needed quality of the assessment. However, with additional resources, the Assessment could perhaps conduct additional research into development procedures and expand its review and field-testing phases. Trade offs in terms of procedures and size of the assessment will continue to be carefully reviewed so that the Assessment can collect the most useful and accurate data possible within the limitations of budgetary capacities.

Selection of Areas To Be Assessed

National Assessment presently conducts assessments in subject areas taught in the schools: reading/literature, writing, mathematics, science, citizenship/social studies, art, music and career development. Several probe assessments, smaller in scope than the subject-area assessments, have also been conducted. Probes allow assessment of areas of immediate, rather than long-term, interest since they are not designed to measure change. They have not been limited to subjects presently taught in the schools.

Probes have included consumer skills, basic life skills, a subset of mathematics skills and adult skills in several areas.

The 1978 legislation authorizing the National Assessment of Educational Progress requires that such an assessment "collect and report at least once every five years data assessing the performance of students at various age or grade levels in each of the areas of reading, writing and mathematics" and that it "conduct special assessments of other educational areas as the need arises" (P.L. 95-561, Sec. 1242). Areas for assessment thus must be selected with regard for these limitations.

Assessments could, however, be structured in ways other than course offerings. For example, a process-oriented assessment that includes tasks drawn from many different subject areas might be developed to determine whether students possess various levels of cognitive skills. Havinghurst's development tasks or Piaget's stages of general intellectual development might be used as a framework. Another potential structure might be an assessment oriented toward problem-solving abilities, measuring children's skills in attacking problems without emphasizing specific content areas. Assessments might focus on a core of basic or functional literacy skills in a number of subject areas and include higher-level skills or additional subject areas as time and money permitted.

These alternatives would result in an assessment quite different from the existing one. In determining the assessment structure, the Assessment designers felt that information on educational progress would be most meaningful if organized in the way that schools are typically organized. Educators -- a major Assessment constituency -- generally are most concerned with teaching a specific content, although they also attend to developmental levels, problem-solving skills, and so forth. Many professional education organizations are structured by content. In addition, Assessment data are intended to provide information about education to the public, and subject-area assessments provide a context that is readily understandable.

Switching to a different organizational scheme might seriously affect measurement of changes over time. Information gathered during the first 11 years of data collection would probably not be relevant if a very different assessment structure were established.

Although individual items could be reassessed, removed from their original context they would probably provide only fragmentary evidence about changes that had occurred.

Changes in the subject areas assessed could be made, although dropping areas would result in some losses in information about differences over time. Areas might be added, particularly if financial support from groups interested in such areas could be obtained. Changes might also be made in existing areas. For example, the areas of citizenship and social studies previously were assessed separately and have now been combined, as have reading and literature. Similar steps might be taken in other areas if it was felt this would provide the desired information in a more cost-effective manner.

The Assessment Policy Committee, a group of 20 members representing business and industry, the general public, chief state school officers, state legislators, school district superintendents, state and local boards of education, governors and teachers, is responsible for making decisions regarding Assessment policies. This committee, using information from NAEP staff and consultants, determines the areas to be assessed. In making its decisions, it weighs the goals of the Assessment, the need to maintain coverage of as many areas as possible, financial resources and the danger that the Assessment will spread itself too thin in attempting to measure too many areas.

Objectives Development

Before assessing a subject area, it is necessary to define the skills, knowledges and attitudes that are to be assessed. To do this, a framework or structure for the learning area is established. Traditionally this framework has been in the nature of objectives, but in recent years content by process matrices have replaced objectives in some areas. For convenience, the developmental procedures will be referred to as objectives development and the framework as objectives, although some areas do not have objectives as such.

If National Assessment results are to be useful to education decision makers across the country, they must represent the myriad interests, points of view and priorities found in American education. To achieve this representation, objectives are developed through a

consensus process. Subject-matter specialists, educators and lay people representing different regions of the country, both sexes, minority groups and different sizes and types of communities create the objectives. As mentioned previously, the goals must be considered important and representative of the discipline by scholars in the field; they should be acceptable to most educators and currently accepted as goals by the majority of schools, and they must be considered important for youth to achieve by thoughtful lay citizens. Objectives serve as guides for the development of assessment exercises; thus, many assessment objectives are stated in such a way that an observable behavior is described.

Objectives in a subject area are reviewed prior to the succeeding assessment in that area and redeveloped if necessary. Usually redevelopment efforts reflect shifts in emphasis in a subject area or new areas of attention rather than resulting in a totally new set of objectives. For example, in science, much of the content covered has remained the same, but scientific processes have received more emphasis in later assessments.

The following discussion of National Assessment procedures focuses upon objectives and exercise development procedures in use from the 1976-77 assessment to the present, since those assessments are most representative of current methods. Table 1 shows the areas assessed between 1976-77 and 1981-82. These areas, with the exception of the 1976-77 adult probe assessments, had been previously assessed, and thus objectives development efforts involved redevelopment and revision of existing objectives.

TABLE 1. Assessment Schedule*

Years	Subjects Assessed	Ages
1976-77	Science	9, 13, 17
	Adult health, energy, reading, science	Young adults 26-35
1977-78	Mathematics Consumer skills	9, 13, 17
1978-79	Writing, art, music Women in Math	9, 13, 17
1979-80	Reading/Literature	9, 13, 17
1980-81	No data collection	
1981-82	Mathematics, citizenship/ social studies (projected)	9, 13, 17

*The NAEP Newsletter (October 1979) provides a complete schedule of areas assessed since 1969.

In recent years, objectives development has been coordinated and monitored by National Assessment staff. For most subject areas, an advisory committee of 5 to 15 members has been convened to provide continuity of personnel during both the objectives and exercise development processes and to advise staff on decisions regarding the subject area when necessary. The advisory committees are selected to represent National Assessment's various constituencies. Members are selected through recommendations from professional education organizations, from previous National Assessment consultants and from other organizations with an interest in education.

Development of objectives has proceeded in slightly different ways in different subject areas. For example, in the most recent science assessment, a key group of

individuals structured the content by process matrix, and other consultants were called in to generate content samples for the matrix cells. For the present reading/literature assessment, groups of consultants assisted staff and the Advisory Committee to consolidate the objectives from the two areas and to amplify the objectives.

The structure of the objectives also varies by subject area. For example, reading/literature consists of four main objectives, each including several subobjectives. These subobjectives have been "amplified" to provide guidance to exercise writers. Science and mathematics are structured through content by process matrices; content samples were developed to indicate content to be covered in science, while a series of questions was used to amplify the mathematics matrix. (For a description of objectives and objectives development in recent assessments, see Art Objectives 1971, 1978; Citizenship and Social Studies Objectives, 1981-82 Assessment, 1980; Mathematics Objectives, Second Assessment, 1978; Music Objectives, Second Assessment, 1980; Reading and Literature Objectives, 1979-80 Assessment, 1980; Science Objectives, Third Assessment, 1979; Writing Objectives, Second Assessment, 1972, 1978.)

Following development, the objectives are reviewed by subject-matter specialists, educators and lay persons. Most reviews have been done in conferences although some, particularly lay reviews, have been conducted by mail.

In describing its approach to objectives development as a consensus approach, National Assessment does not mean that each and every reviewer must agree to everything found in the objectives for a subject area. Objectives or areas considered important by each group of reviewers are added to the objectives, and the objectives embody not the lowest common denominator of objectives that everyone can agree with, but a compendium of many ideas and philosophies. The details of National Assessment objectives development procedures, the names of the consultants used and the objectives themselves are published so that users can judge for themselves the validity of the objectives for their particular purposes.

The objectives developed for subject-area assessments include both what most schools are actually teaching now and what some will probably be teaching in

the future. Although some people have argued that it is unfair to assess topics that are not currently taught in nearly all schools, the Assessment reports on educational achievement, not simply the effects of schooling, and thus includes topics that may not be part of all schools' curricula. A somewhat "future" orientation also allows collection of baseline or "before" data on emerging trends in the school curriculum.

Objectives serve not only as a tool for guiding exercise development but as a definition of what a national consensus of subject experts, educators and lay persons feel is important to an area. Thus, topics or objectives that are difficult to measure are often included in the objectives even though it presently may not be possible for National Assessment to measure them fully.

Potential alternatives to the present approach to objectives development include using a framework other than objectives to structure the assessment, having the development of objectives done by outside contractors, developing objectives through analysis of textbooks, using objectives that have been developed by some other group and changing the size or composition of review groups.

Subject-area assessments might be structured by frameworks other than objectives. For example, items might be addressed to questions about a subject area. In mathematics, such questions were used as the basis of the data reporting scheme. Subject-area consultants and advisory committees will continue to determine the most appropriate structures for the different areas.

Latent traits, such as reading ability or critical thinking, could be assessed. This would represent a departure from National Assessment's present goal of reporting achievements on specific tasks measuring objectives. More investigation of audience interest in information on latent traits and on the best ways to collect such information would be needed.

Different procedures might be followed in developing an assessment framework. In the early days of the Assessment, objectives and exercises were developed by outside contractors. This approach proved unsatisfactory because Assessment staff, who were most familiar with the needs of the Assessment, did not have sufficient control over the developmental process.

Development by an outside contractor might allow use of a contractor with extensive and specific expertise in developing objectives. NAEP staff could then act as critics and reviewers of the objectives produced rather than being intimately involved in the developmental process. Stringent contract specifications and monitoring would be needed. Cost factors would have to be carefully evaluated in light of the advantages gained by contractor development, since past experience has shown objectives development by an outside contractor to be more expensive than "in-house" development.

If objectives were developed through analyzing textbook content, they would probably reflect what is being taught in the nation's schools. Exclusive use of this approach would probably be less expensive than present methods if consultant reviews were discontinued. However, evidence about the extent to which instruction actually follows textbooks is mixed. Textbook analysis also would not allow inclusion of future trends not yet reflected in the texts and would not provide the broad base of support that comes from involving different groups. While exclusive uses of textbook analysis might have drawbacks, some combination of this and other approaches might be effective.

Existing objectives developed by other individuals or organizations could be used. Objectives developers currently draw upon their knowledge of existing objectives in designing objectives for National Assessment. Existing objectives used without review or changes could be challenged as not truly national in scope or not representative of National Assessment's constituencies. Although it might be unwise to simply appropriate existing objectives, such objectives could be relied upon to a greater extent than at the present. If existing objectives were used without changes, some review mechanism would probably be needed to insure that the objectives are sufficiently representative and current. This approach again might be less costly than the present one.

The composition of the development and review groups might be changed, enlarged or reduced. Present policies encourage representation of people on the "cutting edge" in their subject area, which may result in an assessment in advance of actual school practices. This policy might be changed to a more "middle of the road" approach. More emphasis on including users at the federal and state level, such as representatives of federal agencies, might be considered. Extremely large

numbers of reviewers become unwieldy, but enough reviewers should be included to represent various points of view. Assessment staff will continue to monitor the effectiveness of development and review groups to determine procedures that are most useful.

Area Specifications

Objectives outline and define the parameters of the area being assessed but often do not provide guidelines specific enough to structure the work of the item developers. In early assessments, item writers were often instructed to work from the objectives, but this at times resulted in assessments with unbalanced objectives coverage or inadequate content representation. Through the years Assessment staff have become increasingly aware of the need to supply additional guidelines. Area specifications describe the proportion of items to be devoted to different objectives or content areas, type of content or stimulus materials to be included, proportion of items to be administered at more than one age, subobjectives not appropriate to certain age groups, items from a previous assessment to be included and so forth.

In most subject areas, the Advisory Committee alone or working in conjunction with consultants involved in objectives development has recommended to staff the relative importance or weight to be placed on various objectives or cells of the developmental framework.

The detail of subject-area specifications varies with the subject area. Extensive content samples were developed in science and amplified objectives, including samples of expected behaviors, were used in reading. However, in citizenship/social studies, many of the content specifications for the area were developed during the item writing process.

The specifications are generally used as a guide to the development of new items. However, redevelopment of objectives may mean that the objectives for some items to be used to measure change have been dropped or modified. Change exercises must be examined to determine how they will fit into the new objectives scheme in order to specify how many new items are needed for a particular objective. If change items will not fit into the new objective scheme, other plans for reporting them are made. This planning generally begins at the area specification stage.

Although Assessment staff have gained considerable experience in creating area specifications, some issues remain unresolved. Weights for objectives are generally expressed as percentages of the total time available for a subject area. Under this system, not enough time may be available to measure complex areas deemed to be of lesser importance. Usually, the percentage allocation has been adhered to, but variations in this policy might be considered. While general guidelines are developed by the advisory committees, much of the actual specification necessarily takes place in discussions between Assessment staff and item writers. The extent to which decisions should be made by staff or by staff in consultation with the Advisory Board is not always well-defined. Assessment staff will work on establishment of more specific guidelines.

The advisory committees, while constituted to be representative, are relatively small bodies, and bias might enter into their generation of area guidelines. Area specifications might also affect the content validity of an area, since they define content and topics to be included. Reviewers during the item development process should attend to content validity and possible bias, but staff and the advisory committee should be aware of such potential difficulties in designing specifications.

Procedures for developing area specifications are still evolving. It is essential that decisions about an area be made with adequate representation of the groups concerned. Yet, large groups are impractical for making the many and specific decisions crucial to developing guidelines, and those developing guidelines must be fully cognizant of the nuances of the objectives. Presently, an advisory committee seems to be the most appropriate representative body to advise staff.

Other systems for developing area specifications have been suggested. Again, textbook analysis might be used to flesh out the objectives. Task analysis, in which tasks completed by a person considered to have achieved the objectives used as the basis for item development, is another possible strategy. Guidelines might simply be developed by Assessment staff in conjunction with item writers, or the guidelines might be developed solely by the advisory committee without input from other groups. A totally different group of reviewers, distinct from the advisory committee, might be invited to write the specifications to study possible

unrecognized bias in present procedures.

Many critical decisions about area coverage are made in creating area specifications. Budgetary considerations may affect some of these decisions. For example, it may be advisable, in light of restricted funds, to assess some objectives with only a national sample and others with a full sample of subgroups. It may be necessary to decide not to assess objectives that would be too costly to assess in an adequate fashion.

Decisions that impact other phases of the assessment are considered by staff and if necessary referred to the Assessment Policy Committee. Decisions that affect Assessment policy, such as whether or not to assess change in a subject area, rest with the Assessment Policy Committee.

CHAPTER 2

DEVELOPMENT OF ASSESSMENT ITEMS AND BACKGROUND QUESTIONS

Item development includes several phases: exercise writing, field tests and exercise reviews. Following development, items are selected for inclusion in the assessment. Concurrently with item development, background questions, which provide information on respondents' experiences and environments, are developed. Close coordination of these steps is necessary to meet exercise development schedules.

Exercises are expected to meet a number of criteria. Every exercise should be a direct measure of some knowledge, skill or attitude stated in the objectives. An exercise should not be trivial or peripheral to its objective. Exercises should be understandable -- the student must know what he or she is expected to do without confusion due to directions or vocabulary level. Exercises are expected to have a correct answer so that it can be determined whether changes in performance indicate progress or lack of progress toward attaining the subject-area objectives. "Correct" answers in the assessment of attitudes reflect whether or not students meet the goals stated in the learning-area objectives. Some attitudinal items are designed to determine whether students have an opinion on a topic and for these items no particular opinion is designated as correct or incorrect. As mentioned previously, exercises are purposefully designed so that some are easy, some are of medium difficulty and some are hard. Many exercises are designed for more than one age level so that comparisons between age groups can be made.

Writers traditionally have been encouraged to develop exercises for National Assessment in formats most suitable for the task or concept being assessed and not to limit themselves to the multiple-choice, paper-and-pencil mode. Open-ended and individually-administered exercises were frequently used in early assessments. However, such exercises are very

expensive to administer and score, and budgetary limitations have reduced the use of nontraditional exercise formats for the present.

Item Writing

During recent years, items have been written by outside consultants whose work has been coordinated, reviewed and revised by staff. Actual organization of the writing task has varied across subject areas -- in some areas, regional writing centers have been established; in others, local (Denver-area) writers have been primarily relied upon; in others, consultants have come to Denver to write items. Table 2 shows procedures and approximate numbers of item writers used in the various subject areas.

Although the organization of the writing task differed, methods were much the same across subject areas. Writers generally worked in small groups, either in National Assessment conferences or regional centers, and had a chance to review and critique each other's work. Items were also reviewed by National Assessment staff to insure that assessment guidelines were being followed. In citizenship/social studies, writers did not work in groups, but each item writer met individually with staff.

Development through the mail was tried for a small set of music items. This procedure was not particularly effective because it was difficult to give instructions by mail and writers could not interact with each other or staff. Although costs would be lower, mail development does not appear to provide satisfactory, quality items.

TABLE 2. Procedures Used for Developing Items in Various National Assessment Subject Areas

Area	Item Writing Procedures	No. of Writers (Approx.)
Science 1976-77	Item writing conferences of about 6 people for a specific age and topic were held, generally in Denver. Writers included college educators, classroom teachers and curriculum people. Writing centers of 6-8 people each in Boulder and in Denver also produced items.	50
Mathematics 1977-78	5-6 development centers were established at universities with mathematics education departments in various regions of the country. Teachers, graduate students and university faculty worked on items. Each center worked with a different content topic.	50
Writing 1978-79	Exercises were developed in meetings of 3-4 consultants, generally in Denver. Consultants were mostly university faculty specializing in writing and writing education.	10
Music 1978-79	A number of item writing conferences were held in Denver and one conference was held in Denton, Texas. In addition, several writers were identified and asked to send in exercises by mail. Music performance items were developed at 5 writing centers across the country, of 3-8 persons each. Performance items have not yet been assessed.	47
Art 1978-79	Items were developed at conferences held in Denver and elsewhere and through staff	10

	interaction with individual consultants.	
Reading/ Literature 1979-80	Two major development centers were established in Pittsburgh and Minneapolis and center directors recruited writers. Some small item writing conferences also were held.	16
Mathematics 1981-82	Similar to the 1977-78 assessment, 5-6 regional development centers are being used. Conferences on some difficult to measure areas have been held in Denver and St. Louis.	30
Citizenship/ Social Studies 1981-82	Items were developed in large part by local (Denver) consultants working closely with NAEP staff; a number of consultants outside the Denver area were also used. Items were generally reviewed and critiqued in individual meetings with a NAEP staff person.	28

Development monitors try to include exercise writers who represent, so far as possible, minorities, both sexes and a variety of professional experience, including teachers, college educators and curriculum specialists. A danger in using a small number of item writers or a small number of regional centers is that items may reflect only one point of view or education philosophy. Development monitors and the Advisory Committee must be alert to this possibility in selecting item developers to review the items.

Other procedures for writing items might be considered -- for example, using an outside contractor for the entire developmental process including field tryouts and reviews, drawing items from item banks, having national curriculum organizations write items or hiring an internal staff of item writers. In the past, National Assessment has used outside contractors not only to write items, but also to conduct field tests and item reviews. This method did not give Assessment staff sufficient control over the items produced. Large testing contractors at times simply submitted items that they had available and did not develop items

specifically to meet National Assessment objectives. In contracting with outside consultants or writing centers to generate items, National Assessment has assumed more control over the developmental process. An outside contractor for the entire developmental process might reduce NAEP staff time required for item development, but costs and the need to monitor the contractor very closely might be drawbacks.

Items from already existing item banks might be used. However, in order to measure change, it is felt that items must be held secure. Items from item banks would most likely not be secure and results might thus be distorted. Such item banks are regarded as valuable sources for ideas on item generation.

National curriculum organizations might be asked to write items. Although this approach is appealing in that curriculum organizations should be able to mobilize considerable expertise, developmental efforts might reflect the biases of the particular organization and not a consensus of nationwide goals. Curriculum organizations presently serve as monitors and critics, and it might well be important to preserve that function. In addition, such organizations are important users of National Assessment data and materials and it might be awkward if NAEP staff were asked to criticize their work.

Rather than using consultants in a subject area, National Assessment might hire a staff of in-house exercise writers, either temporarily or permanently. This approach could provide greater control over item quality and production schedules. If subject-area specifications are well developed, the same writer should be able to produce items for a number of different content areas. Drawbacks might include inadequate or noninnovative coverage by writers unfamiliar with the subject area and the loss of content validity currently partially achieved by having items developed by people actually working in the subject area. Although the review process should weed out nonvalid items, the resulting time lag required to generate new items might be critical.

Field Testing

Following development, items are field tested to find unforseen difficulties with the items, problems with vocabulary, approximate performance levels and

amount of time required for each exercise. Some reviews may be held prior to field tests, but generally the complete subject-matter, educator and lay reviews follow collection of field-test data.

In field tests, approximately 100 students from a range of socioeconomic backgrounds respond to each exercise. Unlike the regular assessment, intact classrooms are used. Schools used for field tests are not a strictly representative sample but are chosen to represent the different types of students who will be assessed. Items are field tested at each age at which they will be used, since vocabulary problems and timing for each age group need to be determined.

Prior to field tests, items are submitted to the Federal Education Data Acquisition Council (FEDAC) for review for potential invasion of privacy or questions on sensitive topics. Approximately three months must be allowed for preparation of the items and item review; time for these clearance procedures must be built into assessment scheduling.

Field tests -- from sequencing items in booklets, arranging for schools and printing booklets to administering items and at times scoring items -- are conducted by National Assessment staff. Field-test booklets usually are designed to resemble the regular assessment booklets as much as possible. Some requirements are: a mix of easy and difficult items is included in each booklet; items in a booklet must not cue or give away the answer to other items; different types of items and content are included to maintain students' interest. In some subject areas, items covering the same objective have been spread across booklets as in the regular assessments; in others, items measuring the same objective have been packaged together to provide data on validity.

Field-test results are submitted to item analysis, which provides percentages of response for each response alternative or scoring category for an item and point biserial correlations between item and total booklet performance. Item analysis helps to identify items that behave in an unusual manner. These items are then examined by staff and consultants to attempt to determine whether unintended factors are influencing results.

Some possible alternative approaches to field tests include field testing some items during the regular

assessment, using the same set of field-test schools each year or having an outside contractor perform field tests. Items could be field tested during the regular assessment, as items sometimes are in norm-referenced, standardized tests, but such a procedure probably would not be optimally efficient. Although it might be feasible if only a small number of items required field testing, time in the regular assessments is at a premium and adding field-test items would further reduce time available for the assessment. In addition, data on approximately 2,500 respondents -- the number who respond to regular assessment items -- are not needed for field-test purposes. Present field-test procedures often afford an opportunity to interview students about the items, which would not be possible in a large-scale assessment.

Using the same set of schools each year would probably be more convenient and perhaps more efficient. However, this would most likely be an unreasonable imposition on the schools and many would probably refuse to participate. Also, students would become familiar with the type of materials and the assessment situation, thus possibly making results less accurate.

Having an outside contractor perform field tests would be subject to many of the same difficulties found in contracting exercise development -- the need for exact specification of tasks to be performed, the need for contract approval and contract monitoring and the difficulties in communicating with an organization located elsewhere. Administration contractor staff have administered field tests in the past, but this approach did not prove flexible enough to meet the time pressures often encountered.

Modifications to the present system of conducting field tests and using field-test information are constantly being evaluated. For example, collection of information about the amount of time items take is essential since this information is used in making up assessment item booklets, which should take only a class period to administer. National Assessment has experimented with various methods of gathering timing information, none completely satisfactory. Methods of gathering such information need further study.

Use of field-test data to gain some information on item validity is being investigated. Rather than spreading items measuring the same objective across booklets, some items measuring the same objective could

be placed in the same booklet. Correlations between item and total performance for a particular objective would then be possible. Another way to obtain this information would be to have the same students take more than one item booklet. Although high correlations among all items may not be desirable (this would mean all items measuring an objective are measuring exactly the same thing), such correlations could be used as an indication of how items intended to be related performed together. This data could then be used as one piece of information in selecting items.

The Assessment could increase the number of students included in field tests or field test items with different populations of students than presently used. Increasing the number of students would increase costs. For present uses of field-test data such an increase does not appear to be called for. Different populations of students could certainly be used if a specific need for the data arises.

All items developed might not need to be field tested if format and wording of some items are similar to items that have already been tested. This strategy would have to be implemented very carefully to insure that items not tested do not differ in subtle ways from those included in field tests.

Item Reviews

Reviews of exercises serve several functions; to insure exercise accuracy and clarity, to examine content validity and to guard against trivial items or item bias. In many cases, small subject-matter reviews are held prior to field tests to eliminate items that are inaccurate or have obvious deficiencies in content, wording or instructions. Following field tests, items are reviewed by groups of subject-matter specialists, educators and lay people. The reviewers examine items for content validity, correctness with respect to content, importance and potential bias. (For a complete discussion of bias and procedures used to guard against it, see Bias: Psychometric and Social Implications for the National Assessment of Educational Progress, 1980.) In addition to rejecting items or recommending revisions, reviewers also rate items, and their ratings are considered in the item selection process. National Assessment procedures are an attempt to achieve consensus in judgments about items to be included in the assessment.

The number of reviewers and review meetings varies according to the number of items and the nature of the subject area being assessed. Reviews are typically done by conferences of 6 to 8 people. In exercise review, as in objectives development, a broad representation of different types of people from across the country is involved.

Reviewers do not always see all items measuring a particular objective. Thus, while they can make decisions about the content validity of individual items, they sometimes cannot make judgments about the content validity of coverage of the objective. If at all possible, review groups should also be involved in checking objective coverage validity.

An alternative to the present system might be to conduct all item reviews by mail. While this might be less expensive, staff would be left with the task of reconciling opposing points of view, a task that might impose staff biases on the assessment. In addition, opportunities for working on revisions and developing ideas for new exercises through group interaction would be lost.

Another alternative might be to limit the number of reviewers or to eliminate some of the groups now relied upon -- subject-matter specialists, educators or lay persons. Although again costs might be reduced, each of these groups makes a unique contribution to the credibility and validity of the Assessment and loss of any of them might make assessment items less accurate from a content standpoint or less relevant to education concerns.

The types of reviewers called upon might be expanded. Increased review by psychometric experts has been recommended as a step to monitor the measurement qualities of the assessment and to consider measurement concerns regarding particular items or item formats. In some areas, exercise writers have been invited to review conferences in order to clarify the rationales behind their exercises and to discuss suggestions for exercise revisions. This exchange has proved beneficial both for reviewers and writers, and inclusion of item writers will be considered for other areas when funding permits. Administration contractor staff have sometimes been asked to review items for potential administration problems. Their routine inclusion in the review process could be considered.

More extensive checks on content validity than those presently conducted might be useful. In reading/literature, a study was conducted following item reviews in which a different set of reviewers was asked to match items with the objectives they felt most appropriate. Although data from the study are not yet available, staff involved feel that the procedure serves as an additional check on content validity and should be considered for use in other subject areas. Items could be reviewed by more than one group of subject-matter experts or lay people and item ratings compared. The advantages and disadvantages of such an approach have not been fully explored. Costs would have to be weighed carefully against potential benefits.

Item Selection

The actual number of exercises assessed is limited by the time available for an assessment. More items are developed than can be used in an assessment to allow for losses during the review process, and those actually used are selected based on a number of criteria.

Selection is generally done by the Advisory Committee and Assessment staff, drawing heavily upon ratings and comments made by exercise reviewers. People selecting items attempt to design an assessment that includes easy, medium and difficult items; that represents the objectives or matrix proportionally, based upon the weights established during the area specification process; that includes a variety of item formats and stimulus materials; and that includes items that will be appropriate measures of change.

Selection by the Advisory Committee and staff is efficient and helps to insure consistency with the original goals of an assessment. It could be argued that a process as vital as item selection should be conducted by a large group of people. However, the Advisory Committee is felt to be sufficiently representative for the task. Involving new consultants or large groups of people at this stage might well result in chaos or an assessment that varied from the objectives set by the original designers. Reviews prior to the selection process should assure the assessment's validity when a subset of items is selected for use in the actual assessment.

Background Questions

Background questions elicit information about respondents' backgrounds and experiences. This information is used to determine membership in population subgroups whose performance is reported upon and to describe the population of respondents. Some questions are asked of respondents in every assessment; these questions are used to determine membership in groups on which data is always gathered, such as level of parental education and reading material present in the home. These questions are the subject of ongoing research by National Assessment but are not typically a part of the exercise development process. (See Issues in the Analysis and Analysis of Change of National Assessment Data, 1980 for a discussion of National Assessment variables and analysis procedures.)

Other background questions are specific to the different subject areas and are developed during the exercise development process. Some of these questions are included in the item booklets to be answered by respondents; others are included in an instructional program questionnaire that is given to school officials.

Background questions asked of respondents about a subject area generally attempt to gather information about courses taken in school related to the subject area, other experiences with the subject area and related school facilities. For example, in mathematics, students were asked about mathematics courses they had taken, whether they had access to hand-held calculators and whether they had been taught the metric system. In art, students were asked about their experiences in going to museums or galleries, collecting artworks and producing artworks. Such background questions must be included in each different item booklet so that the information gathered will be available for each student and thus can be related to overall achievement.

The instructional program questionnaire given to school officials usually asks for information about the school's programs. These questionnaires provide information about the school, not the individual student who is assessed.

Development of background questions has not followed a standardized procedure -- in some cases, they have been developed by the subject area's Advisory

Committee; in others, by a small group of consultants; in others, by some combination of these. Reviews also have been varied. In some cases, these questions have been included in the item reviews by subject-matter specialists, educators and lay people; in others, they have been reviewed only by the Advisory Committee.

In many studies reported in the literature, background data supplied by students has been found to be reasonably reliable and useful. However, staff have found that students generally cannot provide reliable or detailed information on the nature of their school programs (what particular science curriculum they follow, for example). In addition, background questions take away from the time available for achievement questions. Careful study is needed to determine to what extent background information is useful and to what extent it simply "sounds nice."

National Assessment has found flaws in data acquired through instructional program questionnaires given to school officials. At times, it is difficult to find the appropriate school official to answer the questions. Principals are generally asked but, especially in large school systems, they may not be close enough to programs to provide accurate information. It is sometimes difficult to gather reliable information on courses since course names vary across the country. Officials may fill out questionnaires hurriedly and make mistakes in filling out the forms. In addition, the questionnaires provide information about the school, not the students sampled for the assessment, and there is no way of knowing if the students assessed have actually participated in programs that are available. Thus, it is difficult to use this information except in a very general way to link programs and achievement.

Due to the difficulties encountered in collecting instructional program data, such data have not been extensively reported. A danger in reporting such data is that they appear "clean," but, on close examination, the differences in the way that school officials respond may make the data less reliable than they seem.

There is continuing debate about the role of subject-specific background questions. On the one hand, National Assessment is designed to gather information about student achievement and is not designed to evaluate the school's contribution to that achievement. On the other hand, the Assessment provides an excellent

opportunity to gather information on school programs using a nationwide sample, and many do not want to pass up such an opportunity.

It is important to remember that National Assessment data cannot be used to ascribe causality, although relationships between variables can be described. If this limitation is not well understood, background information may be expected to supply information that cannot in fact be provided. (For a discussion of the limitations in using background information, see Major Informational Needs of National Assessment Audiences and Ways To Enhance the Assessment's Utility in Meeting Those Needs, 1980.)

An alternative to present procedures would be simply to cease collection of subject-related background data. In mathematics, the Advisory Committee and leaders of the National Council of Teachers of Mathematics felt that collection of data on achievement items was more important than data from the instructional program questionnaire and recommended that the questionnaire not be used in the next mathematics assessment. In some other subject areas, advisory committees and curriculum organizations are very interested in data from this questionnaire.

Another possibility would be to collect background information for those who desire it, at a cost. This would not, however, remove difficulties with reliability of data. National Assessment would have to provide very specific guidelines regarding the limitations of the data to help prevent inappropriate uses.

Interest in collection of background data has grown in recent years, in part due to the desire to increase National Assessment's utility to secondary researchers. Expanded collection of background variables brings with it the question of the types of information that should be and should not be collected. Considerable interest in adding policy-related background questions to the Assessment exists. Potentially, the number of such questions is unlimited. It is not yet clear how decisions on adding such questions should be made or whether such questions are within the scope of Assessment responsibilities.

CHAPTER 3

PREPARATION OF ITEM BOOKLETS, ITEM ADMINISTRATION AND SCORING

To judge the usefulness of Assessment data for their particular purposes, Assessment users must be aware of factors that might have affected students' responses to items. The procedures described in this chapter, while not a part of the item development process, impact upon the ways in which the data collected can be interpreted. In designing these procedures, Assessment staff and consultants have attempted to reduce, as far as possible, the effect of extraneous factors on performance so that Assessment reports will provide an accurate picture of what young people know and can do.

Preparation of Item Booklets

Preparation of item booklets involves sorting and then sequencing items into different item booklets. National Assessment uses a matrix sampling approach, and each student responds to only one item booklet. Booklets are designed to be roughly equivalent in terms of item difficulty levels, variety of item types and spread of items measuring different objectives.

Booklet preparation is a very important phase of the exercise development process because of the effect that item context may have upon performance and because standard errors for mean percentages on groups of exercises -- for example, those measuring an objective or subobjective -- become more stable as the items are spread across larger numbers of packages.

National Assessment's objective in booklet preparation, as in other phases of the assessment, is to minimize the effect of external factors. Students mark their answers in the booklets rather than on separate sheets in order to reduce the possibility for confusion, and items are spaced on pages to insure clarity and readability. A mix of difficulty levels, item formats and stimulus types is included in each package so that

students will not become discouraged by too many hard items or bored by a lack of variety. Extensive staff and contractor reviews of item booklets are conducted to guard against errors in preparation. In some subject areas, the advisory committee has reviewed packages; in others, it has not.

Prior to printing the item booklets, exercises must be reviewed for sensitive topics or potential invasion of privacy by the Federal Education Data Acquisition Council (FEDAC). This process requires two to three months and must be planned for in exercise development and assessment scheduling. When exercises are cleared, FEDAC issues a clearance number, which is printed on the booklets.

In assessing changes over time, National Assessment tries to present items in a fashion as nearly identical to the previous assessment as possible. However, the entire item booklet is generally not identical from assessment to assessment since some items in an assessment are released for publication and not reused. The effect of such variation has not been thoroughly investigated. National Assessment attempts to mitigate possible effects by keeping the context for items used to measure change as similar as possible to the context of items used in the previous assessment. In the second assessment of mathematics, all items used to measure change were kept in the same order as in the previous mathematics assessment and placed either at the beginning or the end of a booklet. Context was thus kept similar except for the removal of items released after the first assessment. In reading, very few items were released between the second and third assessments. Items measuring change were readministered in booklets very nearly identical to booklets in the previous reading assessment and were not included in booklets containing new items. This method has the advantage of preserving identical item context; however, care must be taken that change items are spread across a sufficient number of booklets to obtain reliable change data.

Items to be released for publication can be a factor in schemes for arranging items in booklets. Generally, matched sets of exercises are created for subsets of items, and items to be released for publication are selected at random. Often items to be released are not determined until after booklets are prepared. In reading, items to be released and not released were placed in separate booklets so that context for unreleased exercises would be the same in

the next assessment and intact packages of released items would be available for users of Assessment materials. This approach has advantages and disadvantages. A major advantage is that the context for change items remains the same from assessment to assessment. However, concentrating change items into a smaller number of booklets reduces the sample size and care must be taken that a sufficient sample is retained to provide reliable results. Release of intact packages might encourage users of Assessment materials, such as states and local school districts, to simply use the booklets as tests without carefully examining the fit of items within their own objectives.

Because National Assessment is beginning to use booklet scores in the calculation of the achievement-class variable (see Issues in the Analysis and Analysis of Change of National Assessment Data, 1980), it may be necessary to use more care in making booklets equivalent.

Presently, items measuring a particular objective are spread across booklets, making it difficult for researchers, who often wish to work with the sample from only one booklet, to obtain adequate indices of performance. National Assessment might consider designing some booklets so researchers could obtain information on areas of interest to them. Some mechanism would have to be devised for obtaining and selecting researcher input during the booklet preparation process, and care would be needed to insure that the samples needed for the regular assessment would not be adversely affected by such a change in procedures. Another difficulty in designing booklets for research use is the length of time (one to two years) between booklet preparation and availability of data.

Booklets are presently designed to take approximately 45 minutes to administer, including instructions and background questions -- about the length of a typical class period. If it were deemed advisable, the booklets could be lengthened. Longer booklets would require more time from students and might disrupt school schedules, thus perhaps adversely affecting schools' or students' willingness to participate in the assessment.

Current booklet preparation procedures are the result of the combined experience of National Assessment and printing contractor staff over the last 10 years. The procedures are suited to the present design and

goals of the Assessment. If changes are made in other areas of the Assessment, changes in booklet preparation procedures might be needed.

Administration Procedures

Administration procedures affect exercise development insofar as they affect the interpretation that can be made of the results. Critical elements of National Assessment administration procedures include using paced tapes to read instructions and items to students, using specially trained administrators rather than school personnel to administer the items and having students mark answers in the booklets rather than on separate answer sheets.

Assessment procedures are constantly being evaluated to determine whether they can be carried out in a more efficient and cost-effective manner without an unacceptable reduction in data quality. For instance, National Assessment is presently conducting a study on the use of tapes that simply indicate the time remaining rather than read instructions and items to students. This would allow administration of several different booklets in the same time period and thus cut down on administration costs. Data from this study of different tape procedures have not yet been completely analyzed, but preliminary results indicate that changing tape procedures would make a difference in performance levels for some groups. Any decision on such a change would be the responsibility of the Assessment Policy Committee.

National Assessment staff believe that the use of specially trained outside assessment administrators is essential to maintain consistency of procedures across the country. Administration by classroom personnel would leave the Assessment vulnerable to charges that some assessment administrators were giving more help, coaching or encouragement than others.

The effect of using separate answer sheets rather than having students answer in item booklets has not been fully studied by National Assessment. It is known that such a change would be more cost- and time-efficient; however, the potential impact on respondent performance is unknown. A study needs to be devised and implemented to consider the possible impact on achievement, on results for items measuring change not previously assessed with answer sheets and on assessing open-ended exercises.

As technology advances, the Assessment might consider using computer terminals or telephone systems to conduct the assessment. Although such technology might be convenient, careful investigation of the degree of monitoring necessary to insure reliable data would be necessary.

Scoring

Although scoring does not take place until after data collection, scoring procedures are of great concern in the exercise development process, since they can directly affect the reliability and validity of Assessment data. Scoring is currently done by an outside contractor. Scoring procedures are relatively straightforward for multiple-choice items. Consultants check the accuracy of the designated correct answer during the development of the items, and students respond by darkening ovals, which are read by optical scanning machines. These machines have very specific, predetermined error tolerances, and reliability of machine scoring is extremely high.

Scoring open-ended responses calls for some degree of judgment on the part of the scorer. In many cases, the decisions to be made are relatively straightforward, as in the case of scoring answers to computation problems or scoring questions requiring a specific one- or two-word response. However, most open-ended scoring is more complicated. Even questions for which short answers are expected can produce an amazing diversity of responses. In the case of scoring writing essays, decision making becomes even more complex.

Scoring, of course, must be consistent for each response if data are to be given any credibility. Thus, all scorers must use the same scoring criteria, and each scorer must apply the criteria consistently from response to response. When measuring change, National Assessment must be very sure that scoring criteria remain the same from assessment to assessment.

The manner in which an exercise is scored can affect its validity for measuring a particular objective. For this reason, scoring guides are generally developed by exercise writers and are used in making determinations about the exercises. However, such guides are often, at best, educated guesses about the way that students will respond, and complete scoring

guides often are not developed until after field-test data are available. These scoring guides are further refined after data from the actual assessment become available.

Procedures for training scorers are designed to insure that scorers interpret and apply scoring guides in the same manner. (For discussions of procedures used in training scorers, see Procedural Handbook: 1977-78 Mathematics Assessment, 1980; Mullis, 1980.) During scoring, reliability studies are carried out to guarantee that scoring remains consistent. (For a description of procedures in specific areas, see Procedural Handbook: 1977-78 Mathematics Assessment, 1980; Mullis, 1980; Three Assessments of Science, 1969-77: Technical Summary, 1979.) National Assessment has achieved high rates of scorer reliability through extensive scorer training and quality-control measures.

Various procedures have been used to insure scoring reliability in measuring changes between assessments. In science, a 10% sample of responses from the 1972-73 science assessment was rescored with responses from the 1976-77 assessment as a check on scoring consistency. If necessary, all responses are rescored. In mathematics, it was necessary to revise scoring guides from the first assessment to meet second assessment criteria for correct and incorrect responses. All first assessment responses were rescored, using the revised scoring guides, by the same scorers and at about the same time as second assessment responses were scored. Quality-control procedures for checking reliability were applied across both sets of data. In scoring essays in writing, it has been found advisable to score all responses from the previous and present assessments at the same time. It is easy for standards in writing to shift without the scorers' awareness, and essays might then be judged on different criteria in different assessments. In scoring essays, responses for unreleased exercises were not scored or reported immediately following the assessment but were held to be scored with responses from the next assessment. The same procedure was used in scoring art exercises requiring drawing. The extent of scoring or rescoring previous assessment responses will probably remain dependent on circumstances and the nature of the subject being assessed.

National Assessment has made every effort to score responses often regarded as difficult to quantify. It

has pioneered a system of scoring written essays and has done important developmental work in scoring art drawings and music performance.

Other users have occasionally had difficulty replicating National Assessment scoring procedures with sufficient reliability so that results could be compared with National Assessment's. Often such users have not been prepared for the extensive checking procedures required to make scoring reliable. In addition, many of the nuances of scoring judgments are known by NAEP and contractor staff but are not always documented sufficiently for outside use. More careful attention to instruction of outside users is needed if comparisons of results on open-ended or essay items are to be encouraged.

The scoring procedures developed by National Assessment have proved highly reliable and provided a good deal of valuable information. In the past, assessment administrators have at times scored items in the field. This method proved unsatisfactory because many more variations in responses occurred than could be predicted by those designing scoring guides, and each administrator applied his or her own criteria for judgments. Under the present system, scorers are carefully monitored so that if odd responses are encountered, their classification can be determined through joint agreement of NAEP and contractor staff.

National Assessment has found it most efficient to have scoring done by an outside contractor and to have the same contractor do the machine scoring and the open-ended, or hand scoring. Booklets to be scored do not have to be shipped to another location when different scoring methods are needed; in addition, the scoring contractor has a trained staff of scoring personnel that can be called upon and augmented when National Assessment conducts a major scoring effort.

CHAPTER 4

SUMMARY

In examining National Assessment procedures for item development, it is essential to keep National Assessment's purposes in mind. The Assessment is designed to describe performance on specific tasks in various subject areas and to monitor changes in performance over time. National Assessment data serve as an indicator, giving information about the overall condition of American education and pointing out areas in need of investigation, but are not designed to provide evidence about cause-and-effect relationships or information on individual students, school districts or states.

To be useful to those across the country interested in education, assessment data must be of high technical quality. Lacking emphasis on technical quality, it would be all too easy for Assessment audiences to dismiss the data as flawed or inaccurate. Considerable attention has been paid to features seen as necessary to provide data of high quality. For example, the sample design is planned to insure accurate representation of the nation's population at the various ages assessed. Paced tapes and specially trained item administrators are used to provide consistent administrative conditions across the country. Extensive quality-control checks and reliability studies are routinely included as a part of item scoring. Numerous reviewers representing many different viewpoints and constituencies are included in the developmental process to help establish the content validity and freedom from bias of objectives and items.

Although at times it may appear that National Assessment procedures are unnecessarily expensive and time-consuming, those designing the Assessment felt that the methods of objectives and item development used were necessary to provide a broad acceptance and base of support for Assessment data and to insure reliable measurements of change.

As the needs of Assessment users change and as new methodologies emerge in testing and measurement,

assessment procedures may need to change accordingly. As an example, the potential of assessment data for research use is receiving increased attention and procedures are being implemented to make the data more available and more useful to researchers and decision makers.

A number of developmental and administrative procedures are currently under investigation. Recent studies of developmental procedures have indicated that more specific guidelines for procedures to be followed in objectives and exercise development would be useful. Assessment staff have begun to standardize criteria for consultants who review assessment materials for bias, developing specific guidelines for types of people to be included. Although broad representation of reviewers has traditionally been a goal of National Assessment and has in general been achieved, standardized guidelines will help to insure that certain groups are represented.

Increased utilization of computer capacities for documentation and revision of items is presently under consideration. Use of the computer should reduce secretarial time in retyping revised exercises and eliminate some opportunities for error. If the computer system became sufficiently sophisticated, item writers could perhaps make and receive revisions at remote terminal locations. Time and expense needed to implement such a system are being studied.

Users of National Assessment data must evaluate the materials and procedures used to determine whether the data are valid for their purposes; however, documentation of procedures has not been as complete as possible in the past and often has not been standard across subject areas. Efforts to improve documentation procedures are currently under way.

Studies of the effect of paced tapes on performance have shown that the tapes do appear to have a positive effect on performance for some groups. Decisions regarding taping procedures will be carefully evaluated with respect to possible effects on the accuracy of assessment data.

National Assessment staff have identified a number of other areas that might benefit from research. National Assessment could conduct a systematic review to determine optimal procedures for development, investigating to what extent procedures should be standardized across subject areas and when and to what

extent they should be allowed to vary. For example, the number of item reviewers in an area necessarily varies with the type and number of items to be reviewed. However, guidelines could be established for numbers and types of reviewers and more stringent requirements for documenting exceptions to prescribed procedures implemented. Guidelines for development might cover such topics as types of consultants to be included in objectives development, necessary criteria for reviewing items, the role of advisory committees and the creation of area specifications.

Criteria for selecting items for inclusion in the assessment could be refined. In addition, these criteria could be better publicized so users would understand why certain items were included.

Comparative studies of content validity might provide useful information. For example, more than one group might be asked to "weight" objectives and subobjectives for a subject area according to their relative importance and the weightings made by different groups compared. If weightings were consistently similar, it might be determined that such a procedure is not needed. If weightings consistently differed widely, further reviews of developmental procedures might be in order. As a check on item content validity, review groups might be asked to assign items to objectives and their assignment compared with that given by the item writer or another review group. Again, extensive areas of disagreement might indicate a need for further study of procedures.

In addition, content validity of groups of items could be investigated through field-test data. Items could be grouped in booklets so that data on item correlations could be obtained. A study of this type is planned for inclusion in field tests of mathematics items for the 1981-82 assessment.

Format of item booklets should perhaps be systematically studied. Present specifications call for printing only one question on a page unless several questions are related to the same stimulus material. There is a potential for cost savings if more than one item is printed on a page. Research into actual effects of various formats on performance is needed. Similarly, the effect of using separate answer sheets has not been empirically determined. Use of separate answer sheets needs to be studied since their use might result in significant cost savings.

In considering alternatives to present procedures, it must be remembered that a change in one area may affect many other areas, perhaps in unexpected ways. For example, measuring traits rather than objectives might demand entirely new test development and validation procedures. Expansion of exercise reviews or field tests might seriously impact assessment scheduling. Minor changes in item format or item presentation between assessments might affect the reliability of change data. Using released items to measure change would affect results in unknown ways. Thus, the ramifications of various alternatives must be considered not just for their desirability in a particular phase of item development but also in light of their potential impact on the assessment as a whole.

Funding constraints may necessitate different strategies and changes in procedures. The Assessment Policy Committee and staff may need to seriously consider redefining subject areas, organizing assessments on a basis other than subject areas, limiting the scope of subject-area assessments or using less extensive quality-control procedures. At present, funding may not permit complete assessments of mathematics and citizenship/social studies in 1981-82. National Assessment might have to either limit the scope of the assessment in either area or make a choice to assess one or the other. Neither of these alternatives is particularly acceptable to representatives of the subject areas involved. Decisions will have to be made by the Assessment Policy Committee concerning how to best meet the goals of National Assessment and the needs of its constituencies. Any changes will, of course, be extensively studied to determine their potential impact upon measurement of change and on the quality of assessment data.

National Assessment constantly examines methods of item development to insure that they are appropriate to the Assessment's goals, that they guard against bias, that they provide representative coverage of the topics being assessed and that they, as much as possible, reduce sources of errors of interpretation. Efforts to systematize and document existing procedures are ongoing. Research into alternative procedures will be conducted as time and money permit so that the effect of any changes in procedures can be more precisely evaluated.

The designers of National Assessment have attempted to develop procedures for item development and data collection that will make the data as credible and as useful as possible to the Assessment's many audiences. The Assessment will continue to evaluate its goals and conduct research on procedures to determine how it can best provide useful and relevant information on the status of American education.

BIBLIOGRAPHY

Anastasi, Anne. Psychological Testing, 2nd ed. New York: The Macmillan Co., 1961.

Art Objectives, 1974-75 Assessment. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1971. ERIC no. ED 051 255. ISBN 0-89398-001-3.

Beck, Ronald A. "A Critical Review of Content Domain Specification/Item Generation Strategies for Criterion-Referenced Tests." Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 8-12, 1979.

Bias: Psychometric and Social Implications for the National Assessment of Educational Progress, no. 12-IP-54. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1980.

Burton, Nancy W. "Reliability and Validity of Assessment." Unpublished draft. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1978.

Citizenship and Social Studies Objectives, 1981-82 Assessment. Report 13-CS-10. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1980. ISBN 0-89398-076-5.

Cronbach, Lee J. "Test Validation," in Educational Measurement, 2nd ed. R.L. Thorndike, ed. Washington, D.C.: American Council on Education, 1971.

Finley, Carmen J. and Frances S. Berdie. The National Assessment Approach to Exercise Development. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1970.

Issues in the Analysis and Analysis of Change of National Assessment Data, no. 12-IP-57. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1980.

Jensen, Arthur R. Bias in Mental Testing. New York: The Free Press, 1980.

Major Informational Needs of National Assessment Audiences and Ways to Enhance the Assessment's Utility in Meeting Those Needs, no. 12-IP-51. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1980.

Mathematics Objectives, Second Assessment, 1977-78 Assessment. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1978. ERIC no. ED 156 439. ISBN 0-89398-133-8.

Mullis, Ina V.S. Using the Primary Trait System for Evaluating Writing, no. 10-W-51. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1980.

Music Objectives, Second Assessment, Report 10-MU-10, 1978-79 Assessment. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1980. ISBN 0-89398-186-9.

Popham, W. James, ed. Criterion-Referenced Measurement: An Introduction. Englewood Cliffs, N.J.: Educational Technology Publications, 1971.

Procedural Handbook: 1977-78 Mathematics Assessment, Report 09-MA-40. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1980. ISBN 0-89398-143-5.

Public Law 95-561, "Education Amendments of 1978," Nov. 1, 1978.

Reading and Literature Objectives, 1979-80 Assessment, Report 11-RL-10. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1980. ISBN 0-89398-219-9.

Rovinelli, Richard J. and Ronald K. Hambleton. "On the Use of Content Specialists in the Assessment of Criterion-Referenced Test Item Validity." Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 19-23, 1976.

Rubenstein, Sherry Ann and Paula Nassif-Royce. "The Outcomes of Statewide Assessment: Implications for Curriculum Evaluation." Paper presented at the Annual Meeting of the American Educational Research Association, New York, April 4-8, 1977.

Science Objectives, Third Assessment, 1976-77 Assessment. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1979. ERIC no. ED 179 402. ISBN 0-89398-293-8.

Sirotnik, Ken. "An Investigation of the Content Effect in Matrix Sampling," in Journal of Educational Measurement, vol. 7, no. 3, Fall 1970.

Stanley, Julian C. and Kenneth P. Hopkins. Educational and Psychological Measurement and Evaluation. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1972.

Supplement to 1971 Art Objectives, no. 10-A-11, 1978-79 Assessment. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1979. ISBN 0-89398-010-2.

Supplement to the 1973-74 Writing Objectives, no. 10-W-11. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1979. ISBN 0-89398-382-9.

Three Assessments of Science, 1969-77: Technical Summary, Report 08-S-21, 1969-70, 1972-73 and 1976-77 Assessments. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1979. ERIC no. ED 168 901. ISBN 0-89398-297-0.

Womer, Frank B. What Is National Assessment? Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1970.

Writing Objectives, Second Assessment, 1973-74 Assessment. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1972. ERIC no. ED 072 460. ISBN 0-89398-381-0.